Spring 2025

CIS 5450 Final Project

HOME PRICES AND EDUCATION SPENDING ACROSS U.S. STATES

Sukya Williams, Hannah Youssef, Michael Pignatelli

AGENDA

OBJECTIVE AND VALUE PROPOSITION

01 Investigate how state-level education finance indicators and home characteristics jointly influence housing prices.

02 How does public education funding relate to home prices across states?

03 Which features best predict housing costs?

04 Understand how feasible it is to buy a home in different states.

DATASETS USED

Kaggle:
USA
Real Estate

2.2M ROWS

Housing listings with price, square footage, etc.

Kaggle:
US Educational
Finances

2,300 ROWS

Education spending and NAEP test scores by state

DATA CLEANING
+ FEATURE ENGINEERING

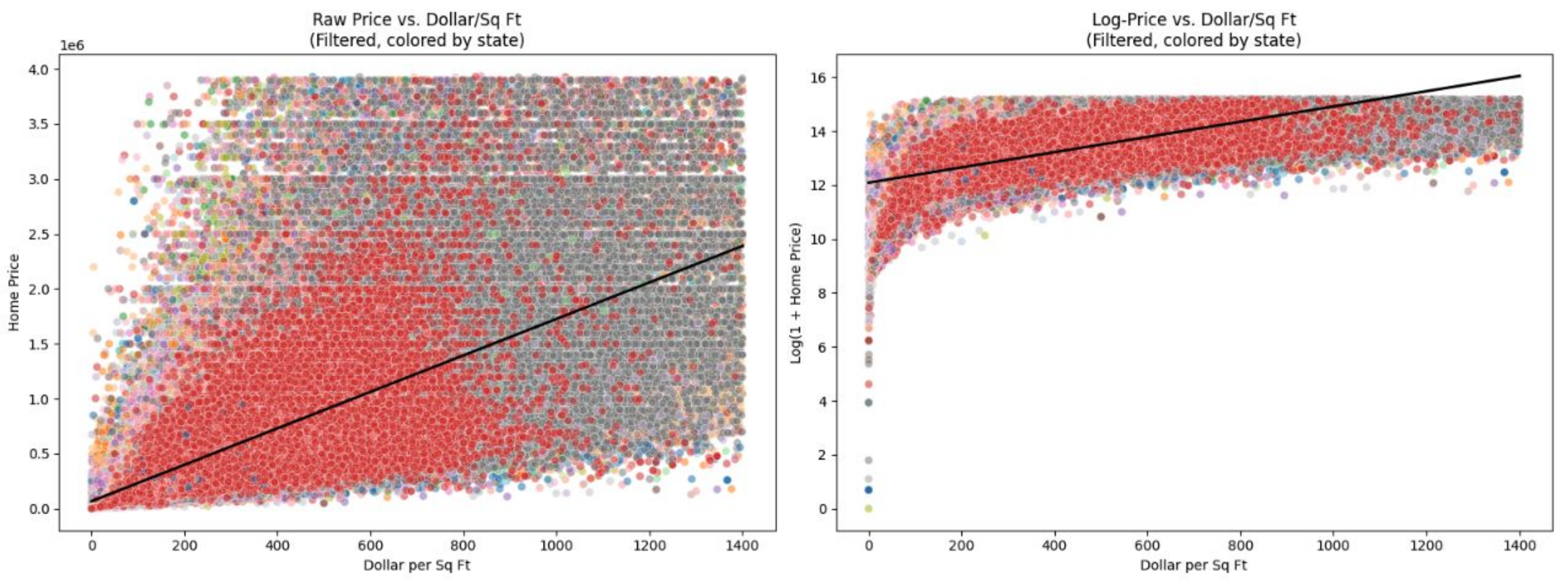
- Filtered by state, dropped missing values
- Engineered new features: dollar_per_sqrt, log_price
- Grouped education data by state
- Joined datasets using Pandas JOIN

~1.6M ROWS

FINAL SIZE AFTER FILTERING

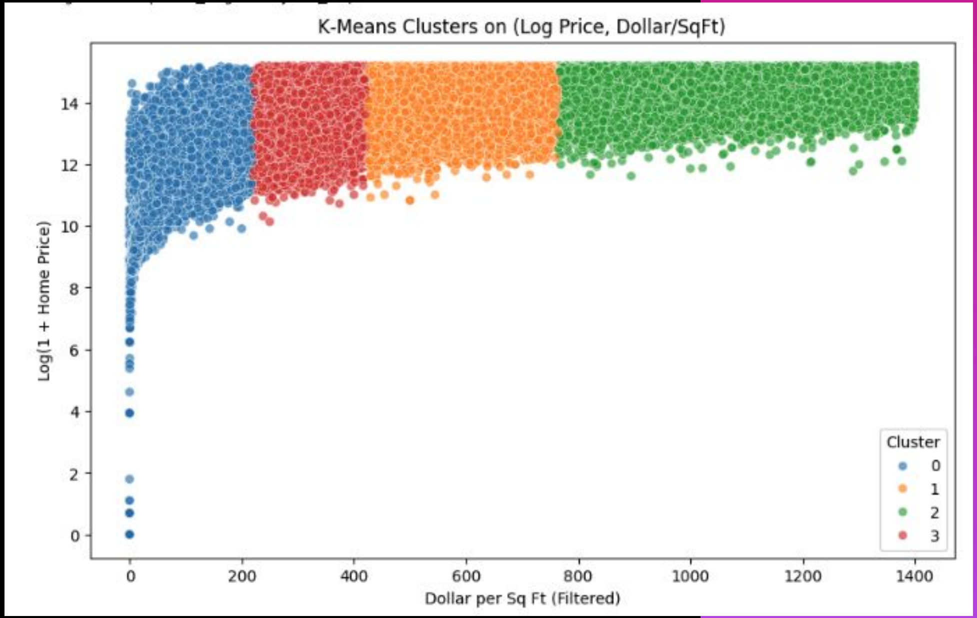
MAJOR LEARNINGS FROM EDA

- There is a linear relationship between dollars per sqft and price.
- The data seems to be clustered by state.



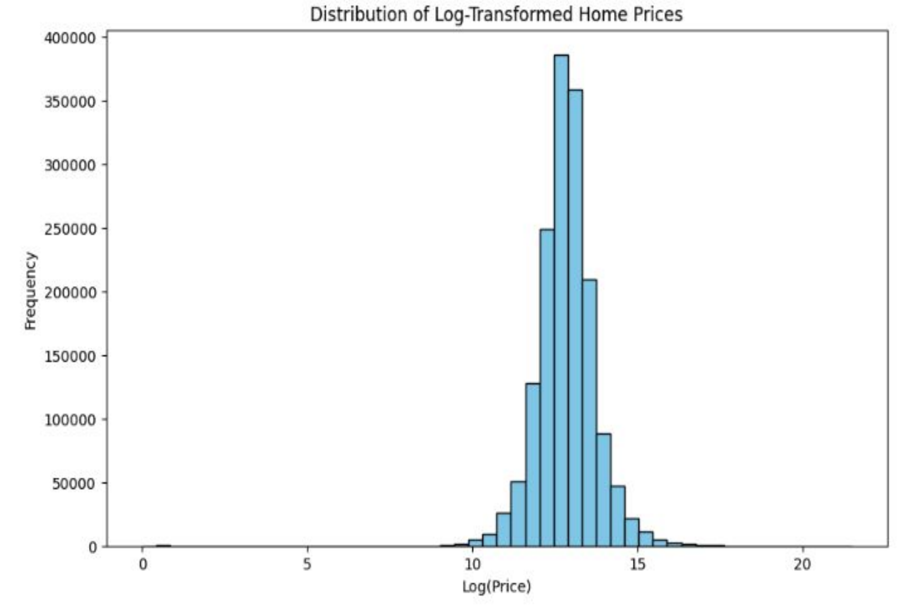
MAJOR LEARNINGS FROM EDA

- All clusters seem to have houses in the same price range, it just seems that some clusters have more points in a specific range.
- Cluster 0: affordable and spacious
- Cluster 1: smaller high-end homes
- Cluster 2: Luxury Homes
- Cluster 3: Moderate price and sqftage



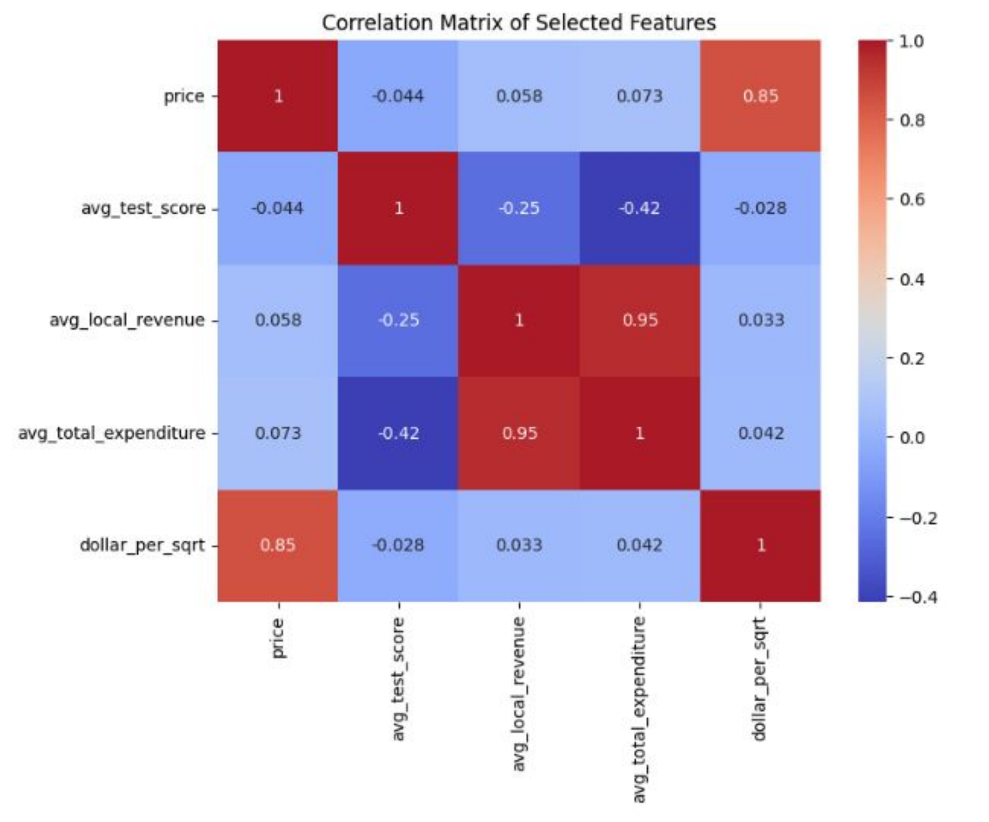
MAJOR LEARNINGS FROM EDA

- The Log transformed graph has a somewhat of a bell curve. This indicates that the original dataset was skewed. The curve is also a bit narrow, so the price range is not ver variable.



MAJOR LEARNINGS FROM EDA

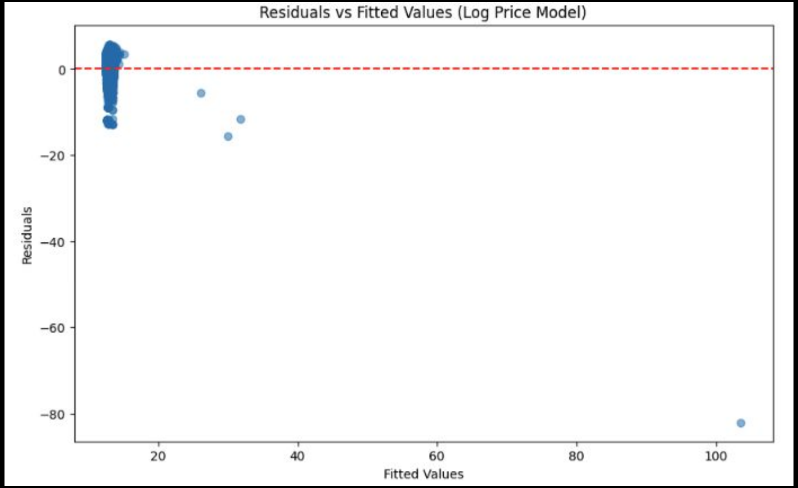
- price is positively correlated with average school total expenditure and local revenue.
- This suggest that areas with high house prices spend more on education and have more non-government spending on education.
- Average test scores have a negative correlation with house price which is counterintuitive!

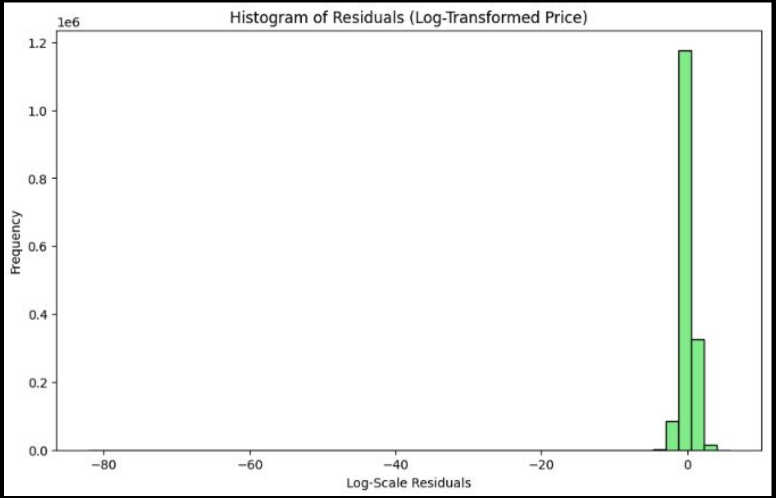


Basic Model

Capturing the features we assumed initially would be sufficient for a decent baseline model on log-price
H0: All coefficients are zero meaning our model has no explanatory power
HA: At least one coefficient is non-zero, meaning that one or more of the predictors have an effect on the log-price

Data Gathered From Basic Model





OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.095
Model:	OLS	Adj. R-squared:	0.095
Method:	Least Squares	F-statistic:	4.222e+04
Date:	Wed, 30 Apr 2025	Prob (F-statistic):	0.00
Time:	13:29:53	Log-Likelihood:	-1.9384e+06
No. Observations:	1602231	AIC:	3.877e+06
Df Residuals:	1602226	BIC:	3.877e+06
Df Model:	4		
Covariance Type:	nonrobust		

Reject the null hypothesis

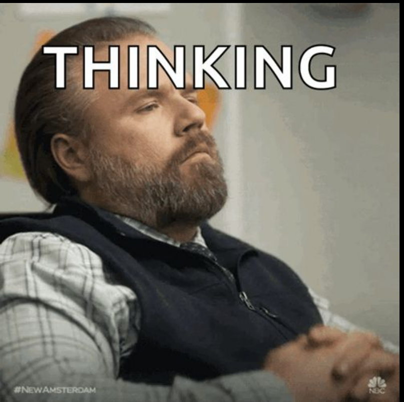
	coef	std err	t	P>|t|	[0.025	0.975]
const	11.0041	0.037	298.974	0.000	10.932	11.076
avg_test_score	0.0068	0.000	45.689	0.000	0.006	0.007
avg_local_revenue	-6.809e-08	2.96e-10	-229.970	0.000	-6.87e-08	-6.75e-08
avg_total_expenditure	3.57e-08	1.25e-10	286.400	0.000	3.55e-08	3.59e-08
dollar_per_sqrt	3.714e-05	3.16e-07	117.606	0.000	3.65e-05	3.78e-05

Does bootstrapping improve the model?

```
Bootstrapped R² Summary:
Mean R²     0.12
Std Dev     0.05
2.5% CI     0.09
97.5% CI    0.23
dtype: float64
```

Mean R^2 = 0.13: indicating that the model explains only 13% of the variance in log_price. Bootstrapping did increase the r2 of our model from 9.2% to 13% but we need a model that captures more of the variance in out log_price.

But…How do I get a model that account for more of the variance in log-price?



Adding More features, addressing multicollinearity and Random Forest Regression

Adding More Features

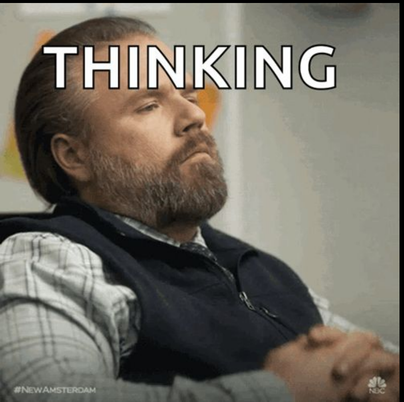
```
const
bed
bath
house_size
avg_test_score
avg_local_revenue
avg_total_expenditure
state_Alabama
state_Alaska
state_Arizona
state_Arkansas
state_California
state_Colorado
state_Connecticut
state_Delaware
state_District of Columbia
state_Florida
state_Georgia
state_Hawaii
state_Idaho
state_Illinois
state_Indiana
state_Iowa
state_Kansas
state_Kentucky
state_Louisiana
state_Maine
state_Maryland
state_Massachusetts
state_Michigan
state_Minnesota
state_Mississippi
state_Missouri
state_Montana
state_Nebraska
state_Nevada
state_New Hampshire
state_New Jersey
state_New Mexico
state_New York
state_North Carolina
state_North Dakota
state_Ohio
state_Oklahoma
```

```
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.422
Model:                            OLS   Adj. R-squared:                  0.422
Method:                 Least Squares   F-statistic:                 2.206e+04
Date:                Wed, 30 Apr 2025   Prob (F-statistic):               0.00
Time:                        13:34:15   Log-Likelihood:            -1.5796e+06
No. Observations:             1602231   AIC:                         3.159e+06
Df Residuals:                 1602177   BIC:                         3.160e+06
Df Model:                          53
Covariance Type:            nonrobust
==============================================================================
```

What’s this?

feature	VIF
const	0.00
bed	2.06
bath	3.02
house_size	2.83
avg_test_score	34118178995231.03
avg_local_revenue	inf
avg_total_expenditure	inf
state_Alabama	4503599627370496.00
state_Alaska	160842843834660.56
state_Arizona	529835250278881.88
state_Arkansas	9007199254740992.00
state_California	inf
state_Colorado	169947155749830.03
state_Connecticut	1286742750677284.50
state_Delaware	195808679450891.12
state_District of Columbia	692861481133922.50
state_Florida	inf
state_Georgia	1286742750677284.50
state_Hawaii	237031559335289.25
state_Idaho	inf
state_Illinois	inf
state_Indiana	4503599627370496.00
state_Iowa	1000799917193443.50
state_Kansas	562949953421312.00
state_Kentucky	321685687669321.12
state_Louisiana	9007199254740992.00
state_Maine	214457125112880.75
state_Maryland	inf
state_Massachusetts	9007199254740992.00
state_Michigan	643371375338642.25
state_Minnesota	529835250278881.88
state_Mississippi	111199990799271.50
state_Missouri	200159983438688.72
state_Montana	204709073971386.19
state_Nebraska	16899060515461.52
state_Nevada	inf
state_New Hampshire	643371375338642.25
state_New Jersey	3002399751580330.50
state_New Mexico	900719925474099.25
state_New York	9007199254740992.00
state_North Carolina	3002399751580330.50
state_North Dakota	391617358901782.25
state_Ohio	2251799813685248.00
state_Oklahoma	562949953421312.00
state_Oregon	1801439850948198.50
state_Pennsylvania	9007199254740992.00
state_Rhode Island	600479950316066.12
state_South Carolina	237031559335289.25
state_South Dakota	35184372088832.00
state_Tennessee	163767259177108.94
state_Texas	1801439850948198.50
state_Utah	1801439850948198.50
state_Vermont	12025633183899.86

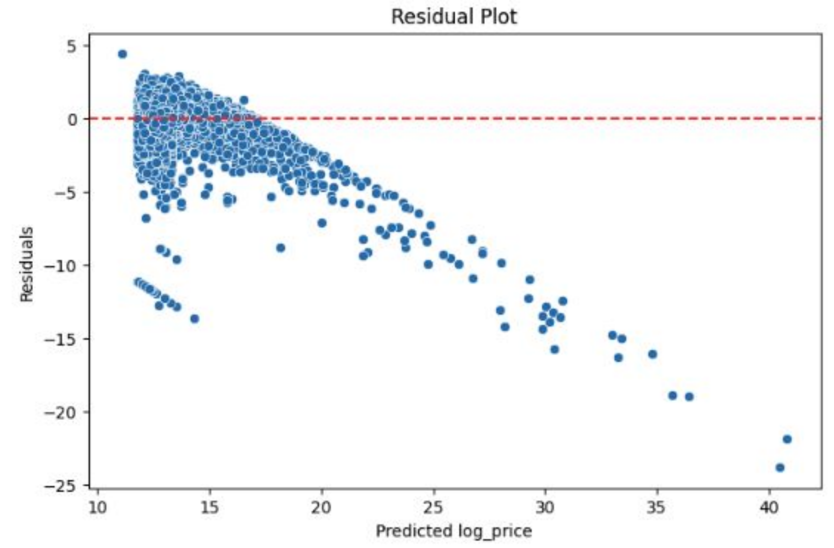
But…How Did We Address Multicollinearity in our Model?



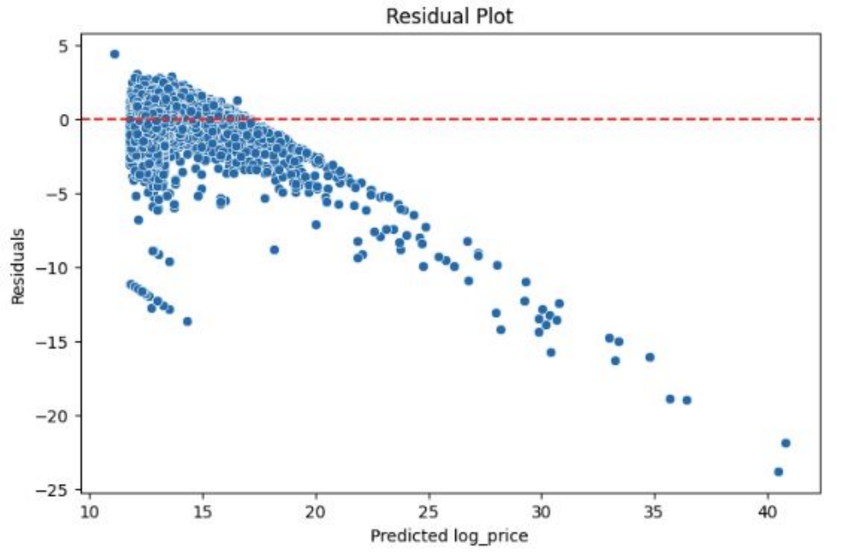
Using Ridge Regression

Ridge Regression

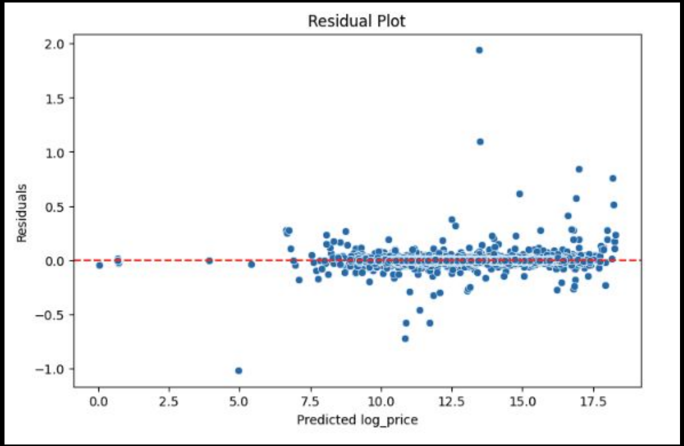
```
Ridge Model Without Tunning................................

The train score for ridge model is 0.47023714822184226
The test score for ridge model is 0.5009415928793077
R^2 Score on Test Set: 0.5009
```



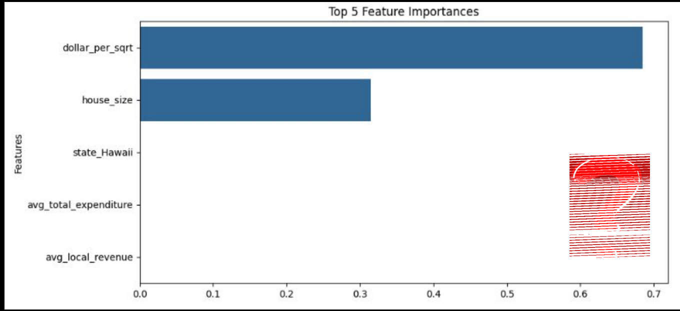

```
Best alpha from cross-validation: 0.1

Ridge Model with Tuning .................................

Train R^2 score: 0.4702
Test R^2 score : 0.5009
```



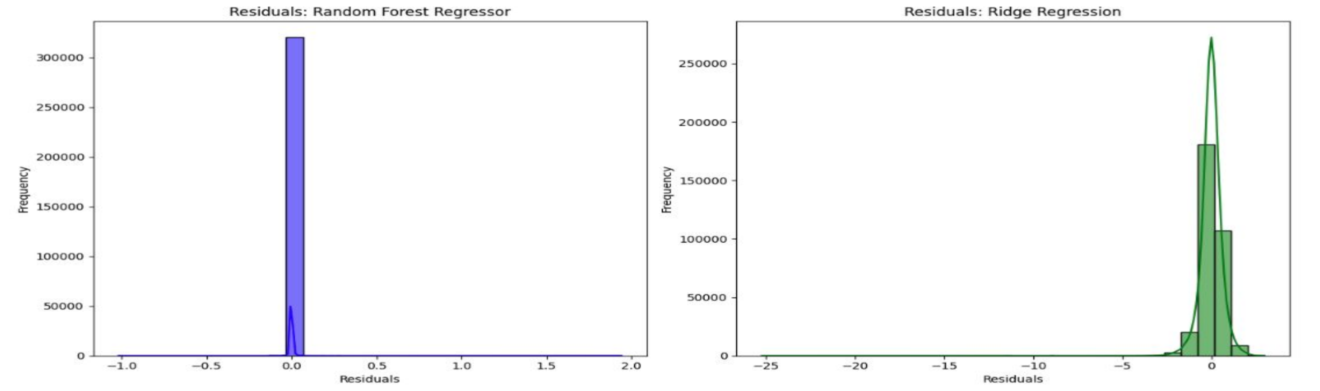
Random Forest Regression Results



R2: 0.9999371732393809
RMSE: 0.006760915929494215



Overfitting
???



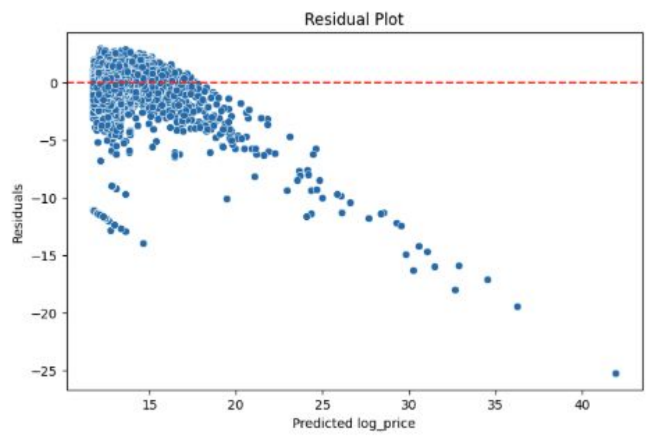
Overfitting

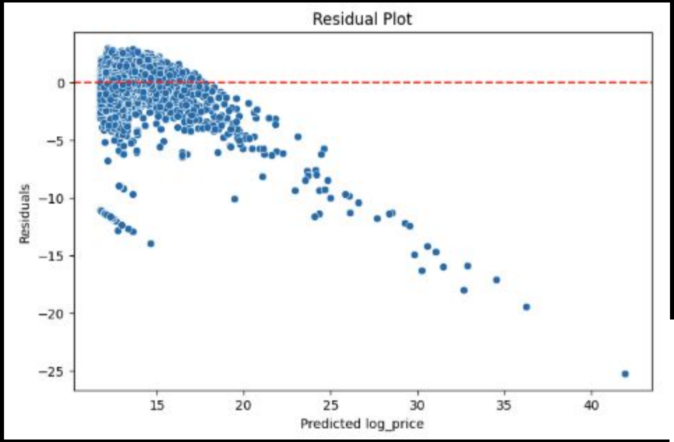
feature	VIF
const	0.00
bed	2.14
bath	3.11
house_size	[illegible]
dollar_per_sqrt	[illegible]
avg_test_score	75059993789508.27
avg_local_revenue	9007199254740992.00
avg_total_expenditure	inf
state_Alabama	1801439850948198.50
state_Alaska	5762763438733.84
state_Arizona	64337137533842.25
state_Arkansas	107228562556440.38
state_California	inf
state_Colorado	180143985094819.84
state_Connecticut	4503599627370496.00
state_Delaware	3002399751580330.50
state_District of Columbia	128674275067728.45

After Correcting Overfitting

OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.422
Model:	OLS	Adj. R-squared:	0.422
Method:	Least Squares	F-statistic:	2.206e+04
Date:	Wed, 30 Apr 2025	Prob (F-statistic):	0.00
Time:	13:34:15	Log-Likelihood:	-1.5796e+06
No. Observations:	1602231	AIC:	3.159e+06
Df Residuals:	1602177	BIC:	3.160e+06
Df Model:	53		
Covariance Type:	nonrobust		



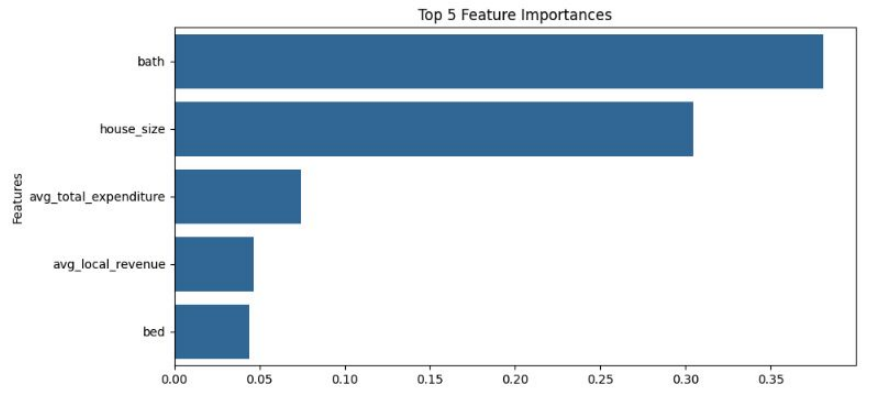


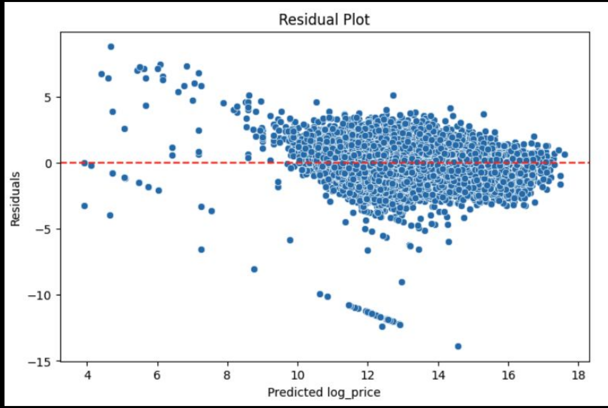

```
Best alpha from cross-validation: 100.0

Ridge Model with Tuning ....................................

Train R^2 score: 0.4421
Test R^2 score : 0.4830
```

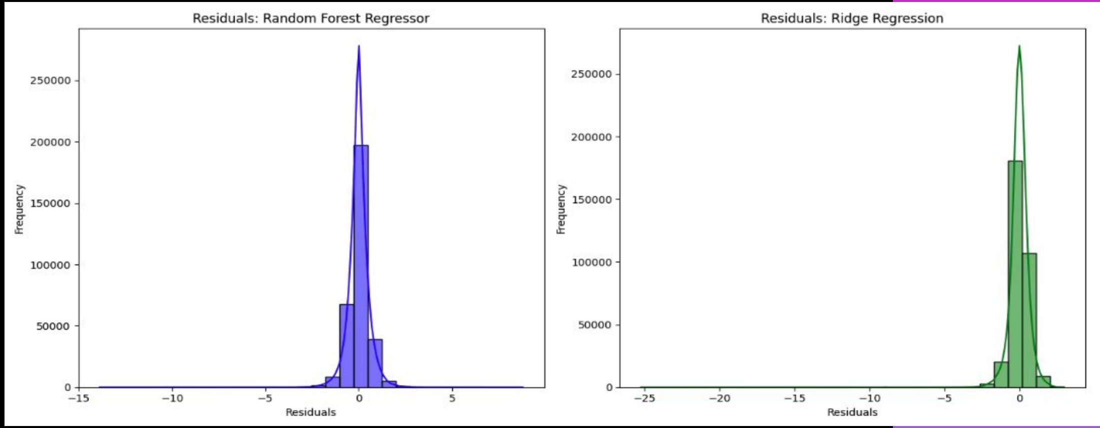
```
Ridge Model Without Tunning................................

The train score for ridge model is 0.44214543774457626
The test score for ridge model is 0.48299199212362887
R^2 Score on Test Set: 0.4830
```

After Correcting Overfitting

R2: 0.5735841761230367
RMSE: 0.5569931548714439







HYPOTHESIS TEST RESULTS

- **OLS Regression & Significance:** using log_price as target, we found that the majority of our predictors had p-values < 0.001 (reject Null Hypothesis)
- **Bootstrapped Confidence Intervals:** we ran over 100 iterations to simulate sampling distributions which improved the R-squared of our baseline model and coefficients were not equal to zero.
- **Interpretation:** *The educational factors that we looked at are statistically significant and our models do have some explanation power.*

IMPLICATIONS AND INSIGHTS

- **Education Spending and Housing Prices Are Statistically Linked:** hypothesis testing supports that local revenue and test scores have a significant relationship with housing prices at the state level.
- **State-level Variation is significant:** some states consistently clustered at higher or lower price levels, suggesting that other features play a large role beyond just home or school-level factors.
- **Home Size and Price per Square Foot Are Dominant Predictors:** basic property features still drive home value even when controlling for education factors.

CHALLENGES AND LIMITATIONS

- **Multicollinearity:** Strong correlations between features like local revenue and expenditure required using Ridge Regression and feature dropping to stabilize the model.
- **Imbalanced Data Representation:** Some states had significantly more observations than others, leading to biased model training.
- **Regression Assumption Violation:** OLS regression assumes linearity and independence of errors, which did not fully hold in our dataset.

POTENTIAL FUTURE WORK

- **Incorporate External Economic Factors:** Add variables such as unemployment rates, interest rates, or median income to see how economic conditions influence housing.
- **Explore at a finer scale:** Use district-level or zip code-level housing and education data to better capture local variations.
- **Interactive Visualization:** Use Plotly or Folium to allow users to explore education/housing relationships by state or region.
- Factor in the use of the city with features used to see if that will improve the R-square of our strongest model, RFR or if this be an important feature.

THANK YOU